2020

University of Massachusetts Amherst Response to Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research

Thea P. Atwood
University of Massachusetts Amherst, tpatwood@umass.edu

University of Massachusetts Amherst Response to Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research

Responder: Thea Atwood, University of Massachusetts Amherst
Response: Discipline Neutral
Role: Data Services Librarian, R1 Public research and land-grant university

Dear Chief of Staff Bonyun, Dr. Nichols, and the Subcommittee on Open Science,

The University of Massachusetts Amherst (UMass Amherst) is pleased to offer its comments on the "Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research". UMass Amherst is a public research and land-grant university with R1: Doctoral Universities -- Very high research activity classification. A generalist data repository, designed to capture the scholarly output of the organization, is managed by the University Libraries.

As described in the RFC, the OSTP seeks comments on key characteristics for data repositories for data resulting from Federally funded research. Specifically, the OSTP is seeking public comment on: I. The proposed use and application of the desirable characteristics; II. The appropriateness of the "Desirable Characteristics for All Data Repositories" (Section I) for data repositories that would store and provide access to data resulting from Federally-supported research; III. Appropriateness of the characteristics listed in the "Additional Considerations for Repositories Storing Human Data (even if de-identified)" (Section II) delineated for repositories maintaining data generated from human samples or specimens; IV. Considerations for any other repository characteristics which should be included to address the managing and sharing of unique data types (e.g., special or rare datasets); V. The ability of existing repositories to meet the desirable characteristics; VI. Consistency of the desirable characteristics with widely used criteria or certification schemes for certifying data repositories; and VII. Any other topic which may be relevant for Federal agencies to consider in developing desirable characteristics for data repositories.

We write on all enumerated topics.

I. The proposed use and application of the desirable characteristics

The term "preservation:" Digital preservation has particular and rigorous requirements that, based on current infrastructure available for universities and others to adopt, may not be sustainable or achievable. Further, there is confusion around business models of preservation services, worries about the cost of preservation-level storage, unsolved legacy issues, and more.[1] Alternative options to describe the concept of making data available for the long term include providing a precise definition for "long-term preservation," including minimum number of years, processes necessary to ensure data integrity, stipulations and workflows for format and hardware migration, etc., or using a different phrasing, like "long-term accessibility," again, with a definition of minimum requirements. Further complicating this issue is that disciplines and other stakeholders have different ideas of what "long-term" means, so some guidance for faculty and repository managers would be welcome.

This document stipulates that "these characteristics are not intended to be an exhaustive set of design features for data repositories" which is admirable -- it helps maintain some flexibility. This paragraph continues: "Federal agencies would not plan to use these characteristics to assess, evaluate, or certify the acceptability of a specific data repository, unless otherwise specified for a particular agency program, initiative, or funding opportunity" which reads as a statement that contradicts itself -- "we won't use this to assess repositories, except when we will." And while there is emphasis that the characteristics should guide Federally funded investigators, this might create inequity in the development and use of repositories -- if there is a seal of approval provided by Federal agencies, would researchers be dissuaded from using an otherwise appropriate repository? Further, organizations unable to meet the certification requirements (because of limited resources, administration not understanding the need to pursue a certification, lack of awareness, or other bureaucratic holdups and misinformation), will be negatively impacted.

Finally, because researchers frequently lack the expertise or the training in using digital repositories, we recommend the final report include a section encouraging consultations with local experts and online educational materials. This includes working with local librarians, data curators, security officers, privacy officers, cybersecurity specialists, and other experts. Federally coordinated data repositories need to be encouraged to include guidance on using the data repository. Critically, researchers

[1] See Rieger (2018). The state of digital preservation in 2018: A snapshot of challenges and gaps. ITHAKA S+R. https://doi.org/10.18665/sr.310626

need guidance on preparing data for deposit, as some of the steps for preparing shareable data need to occur at the beginning of a project -- for example, writing a consent form to appropriately secure consent to share human subjects data digitally.

II. The appropriateness of the "Desirable Characteristics for All Data Repositories" (Section I) for data repositories that would store and provide access to data resulting from Federally-supported research

A. Persistent Unique Identifiers

Persistent, Unique Identifier (PUIDs or PIDs) are critical for the success of data citation, access, and reuse. Downstream effects of PIDs should be explicitly stated to help researchers understand the importance of assigning a PID.

We recommend that the final report require support for digital harvesting technologies, like APIs.

The recommendation would be strongest if it includes a recommendation for use of centrally registered DOIs, and exclude accession numbers. DOIs facilitate tracking use of datasets by working off of robust and standardized infrastructure that locally assigned accession numbers lack.

B. Long-term sustainability

Renaming this section "business model and long-term sustainability," would help clarify its purpose, if indeed this is a section aimed at the business model of the repository.

If so, this section would benefit from providing guidance on what business models look like, or a minimum-viable business model. For example, the University of Massachusetts Amherst uses a hosted repository service, and has one full-time librarian dedicated to managing the repository, as well as some support staff. What *specific* business model characteristics should be included on an "about" page?

Such characteristics might include stipulations on "sunsetting" a repository -- this might be a challenging task to complete, but is an important thought exercise. Again, guidance on this sub-characteristic should be provided.

C. Metadata

Metadata is a critical component in helping others understand a dataset, as well as for findability, and reuse of data. Like with PUIDs, it should be made clear to faculty the

downstream benefit of well-applied metadata.

Furthermore, the word "sufficient" requires additional explanation, as metadata standards vary in complexity and breadth. Researchers will have a very different definition of 'sufficient' as compared to librarians or data curators -- in part a reflection of each group's expertise.

As with other characteristics, the section will benefit from well-defined terminology, with links to resources for further education provided. Using discipline-neutral, jargon-free language in this section is a high priority.

Providing resources on how to evaluate metadata options would greatly improve the usability and reach of this document. Adoption of any metadata standard for a discipline is very poor. Little – if any – guidance on evaluating metadata options is available, further demonstrating a need for providing such guidance.

D. Curation & Quality Assurance

This section would benefit from definitions of 'curation' and 'quality assurance.'

Further, we are concerned about the ability of repositories to actually meet these requirements. More detail is in section V.

E. Access and F. Free & Easy to Access and Reuse

The distinction between "Access" and "Free & Easy to Access and Reuse" is subtle. These two categories should be combined.

"Free" may also be a term that requires some expanding -- some repositories charge a fee on data deposit[2] -- would researchers be dissuaded from using repositories that they have to pay to deposit data? Funding structures for data repositories is still very immature.

Consider including guidance on licensing data, which will explicitly state conditions for use and reuse.

F. Free & Easy to Access and Reuse

Described above.

[2] E.g., Dryad charges $120 as a base fee for data deposit: https://datadryad.org/stash/publishing_charges

G. Reuse

This section would benefit from being renamed "use and reuse", as using a quantifier like download counts doesn't necessarily demonstrate reuse, but it could demonstrate use (e.g., using a dataset to teach). Further, the infrastructure for tracking and counting data citations, repository page views, and downloads is still immature, and not all repositories will have the ability to buy-in to a program or widget or module with this capability.

The "Use and Reuse" section should be more explicitly defined to include what infrastructure a repository will need to meet this requirement.

H. Secure

It is clear what standards repositories should try to adhere to, but it is unclear how a researcher would assess how this is rolled out in a repository. Will researchers worry that a repository is non-compliant if it uses a different security standard than the two suggested here? Will researchers be deterred from using an otherwise appropriate repository? Will researchers feel anxious about depositing data in repositories that do not seem to require security controls, like those that only publish data that can be made openly available?

If this section refers to *user* data (as in password and login information), this should be made clear.

I. Privacy

This section would benefit from examples of what is meant by "administrative, technical, and physical safeguards," as well as the "applicable privacy, risk management, and continuous monitoring requirements."

This section will not be clear to researchers, who won't have the language to assess a compliant repository. This section would benefit from pointing to information security and cyberinfrastructure officers, or data librarians to help provide more robust guidance.

J. Common Format

The metadata characteristic (c) would benefit from a statement on the capabilities of a repository to export metadata to a common format.

Researchers would benefit from links out resources explaining the different types of

non-proprietary formats, and for what format of data (GIS, images, video, text documents, etc.).

The document can help improve other researchers' ability to reuse data by explaining that some data formats, while non-proprietary, do not facilitate reuse (e.g., PDFs in general, tables stored as images).

K. Provenance

As stated, this characteristic sounds like back-end functionality for the repository, and not all repository platforms have the capacity to log file differences at the bit level.

This section would benefit from clarity -- Who is responsible for gathering this information -- the researcher, or the repository manager? If it is the researcher, they will need guidance to locate and capture logfiles at the beginning of their research process, or guidance on workflows and programs that capture this information on their behalf. What level of detail is necessary?

Further, "beginning with creation/upload of the dataset," are two entirely different concepts, at two entirely different points in time, and would require two different workflows -- please clarify what is meant here.

III. Appropriateness of the characteristics listed in the "Additional Considerations for Repositories Storing Human Data (even if de-identified)" (Section II) delineated for repositories maintaining data generated from human samples or specimens

This section seems to note that even de-identified data will need to be stored in a repository that includes the controls outlined here. If so, this will be a significant change in current expectations, and may place new obligations on faculty working with human subjects data. For example, faculty may need to secure funding to store their data in a secure repository, like ICPSR. This may also not be in line with what funders are asking -- so some congruence between funders and the OSTP with regard to human subjects data will be necessary.

Furthermore, what will be done with de-identified data that resides in repositories without these considerations? For example, Dryad accepts de-identified human subjects data, and does not have gatekeeping in place.

More generally, we do not submit any specific comments on the considerations for repositories storing human subjects data, but would again reiterate that most researchers do not have a background in cybersecurity or data curation, and will need more guidance than what is provided. As above, we recommend providing suggestions for where researchers can find help for evaluating repositories and characteristics they are not familiar with.

Finally, and noting this is beyond the scope of this document, it may be necessary for consent forms to explicitly state that data will be made available in aggregate into perpetuity, so that subjects can consent to this type of data sharing. Further, if raw subject data needs to be kept into perpetuity, this should be noted in the consent form. The OSTP should consider coordinating with relevant funding agencies, and how consent to share will be executed by the IRB and other relevant offices at organizations that receive grants.

IV. Considerations for any other repository characteristics which should be included to address the managing and sharing of unique data types (e.g., special or rare datasets)

Some datasets include information that should not be released publicly -- including locations of protected species of plants and animals, or sites that have religious or cultural importance. Of particular concern are images of these items -- images often have GIS coordinates embedded in the metadata. These sharing considerations are not represented anywhere in the guidance, and should either be incorporated in Section II, or in a new Section III relating to special cases.

V. The ability of existing repositories to meet the desirable characteristics

There is a concern that this will become an unfunded mandate for repositories to meet. In particular, the "Data Curation and Quality Assurance" guideline and the requirement for long-term preservation will be a challenge for anyone but the most robustly staffed to provide. These are laudable goals, but fully realized, will be out-of-reach for many. "Expert curation" requires a great deal of staff and time -- an estimate given at the recent Accelerating Public Access to Research Data Summit held in Washington, DC[3], stated that for 150 projects, three individuals are needed. This would be compounded if different federal agencies take different approaches to data curation and quality assurance

[3] February 19, 20, & 21, 2020 - AAU/APLU Workshop & Summit on Accelerating Public Access to Research Data: https://www.aplu.org/projects-and-initiatives/research-science-and-technology/public-access/

VI. Consistency of the desirable characteristics with widely used criteria or certification schemes for certifying data repositories

The guidelines provided seem to be in-step with other used criteria.

VII. Any other topic which may be relevant for Federal agencies to consider in developing desirable characteristics for data repositories.

Consider using the more broadly applicable “open scholarship” over “open science” when referring to open science as a methodology. This can also help expand our definition of scholarly output to include not just data, but curricular materials, teaching outputs, gray literature, primary documents, null data, and more. By using “open scholarship” as the default descriptor, we are inclusionary towards all disciplines and types of data.

Consider creating a template for repositories to use -- something standardized to help researchers quickly assess how their selected repository fits with the recommendations laid out above. This would also be useful for repository managers and institutions to understand what resources they will need to commit to ensure they meet federal guidelines.

Use non-jargon, discipline-neutral language throughout the guide.

Two more complex issues related to data sharing, copyright and intellectual property (IP), are missing from the discussion. Copyright is challenging because it does not uniformly apply to all data types -- e.g., taping an interview of a participant for a language study is considered copyrightable, and ownership can be shared between the researcher and the participant if explicitly stated as such. How IP relates to data is largely unexplored, and is regularly cited as an anxiety for not sharing data. Resources on licensing, copyright, and IP should be included, as these are incredibly confusing topics for researchers, and there is little guidance in existence.

Remind researchers that many institutions have both research data and cyberinfrastructure professionals to answer their questions. Many organizations have an institutional repository to deposit data when a disciplinary repository isn’t available. Offering both of these points of guidance will help reduce confusion, spread accurate information, and improve compliance. However, small grantees, e.g., recipients of SBIR funding are unlikely to have such infrastructure or resources; special provisions may be required for this class of federal contractors and grantees.

Finally, Academia is very late in treating data as an asset. For-profit organizations and publishers see this gap in our services. Companies have already figured out how to mine data for profit, and have access to resources and their own proprietary datasets to create shiny solutions for campuses and researchers. Academia is thus subject to for-profit initiatives that will leave us again in the same place we are with publications -- where publishers own the taxpayer-funded scholarly content. If a high degree of data curation and quality assurance is a requirement that receives no additional funding or support from funding agencies, we will be destined to again rely on the deep pockets of for-profit publishers and companies (and publishers, with their extensive profit-margins, will easily be able to buy up third-party options, if they aren't significant backers already) to meet this requirement. This will extinguish any capability we have of truly meeting the promise of scholarship -- to better our lives, the lives of others, and humanity as a whole.

Conclusion

We thank you for the opportunity to comment on this important matter, and hope that our comments prove helpful. Please feel free to contact Thea Atwood (tpatwood@umass.edu) about our comments.

Sincerely,

Thea Atwood, MSLIS
Data Services Librarian